Enhancing Policy Analysis with NLP: A Reproducible Approach to Incentive Classification

M.A. Waskow and **John P. McCrae**
Data Science Institute
University of Galway
Insight Research Ireland Centre for Data Analytics

Abstract

Over the past few decades, political scientists have increasingly used Natural Language Processing (NLP) methods in their research. Within the subdomain of public administration, there remain further opportunities for the use of NLP in the task of policy analysis. The potential of a machine learning pipeline to identify sentences containing incentives has been demonstrated for the Spanish-language forestry policies of five Latin American countries, but the study was not reproducible due to a lack of model and data availability. This paper seeks to validate the existing pipeline of policy collection, sentence labelling, fine-tuning, and incentive classification by replicating it in a new context and achieving comparable performance, as well as to publish all relevant data and training information to ensure reproducibility. In the domains of a different language and geopolitical system, namely English-language Irish forestry policies, this implementation demonstrates the pipeline's transferability by achieving mean overall F1 scores of 88.3 for binary classification and 96.8 for multiclass classification with our best models. The contributions of this paper are twofold: the validation of an existing pipeline by replicating it in new geopolitical and linguistic domains, and the creation of a novel open dataset of Irish forestry policy sentences labelled for incentive classification.

1 Introduction

Political and social scientists have been utilizing Natural Language Processing (NLP) methods in their research for decades in order to better analyse text as data (Laver et al., 2003; Grimmer and Stewart, 2013). NLP-based analyses of political communication texts — including speeches, manifestos, news articles, and tweets — have provided unique insights into political parties, propaganda, and public opinion. While the field of public policy has been slower in implementing NLP, researchers are beginning to explore its potential applications in the area of policy analysis.

A core task of policy analysis is the development of reports to inform the creation or maintenance of policies (i.e. official proposals of actions to be taken by a government or institution) by presenting evidence-based solutions to a given problem. The research behind these reports can help prevent policy failure by combatting the inappropriate solutions and ambiguity that threaten successful policy implementation, but the thoroughness of policy analysis requires a high time cost that an analyst cannot always afford. NLP can help to reduce this bottleneck of policy creation, making the process more efficient and leading to earlier real-world changes.

Firebanks-Quevedo et al. (2022) demonstrated that their NLP pipeline could help streamline key tasks of policy analysis: finding policies about a given topic and identifying the policy instruments used within them. Policy instruments are methods of government intervention to achieve certain outcomes, like laws (regulation), grants (economic action), or propaganda (communication); this project focused specifically on economic instruments called incentives, which encourage actions and behaviours through financial rewards or pressures like subsidies or taxes (Badie et al., 2011). The pipeline consists of scraping policies from the websites of five Latin American governments, extracting and labelling their sentences for presence and type of policy incentive, then training both a binary and a multiclass classifier on the labelled dataset. The original dataset and models are not available, however, limiting the reproducibility as well as the implementation of their work.

Despite Dodge et al. (2019) offering a structured checklist to address the reproducibility crisis in NLP, many projects do not make core information relating to their research and the production of their results available. This challenges not only the cred-

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 74–85, 2025

ibility of the conclusions, but also the feasibility of downstream impacts of their findings. Magnusson et al. (2023) has shown that even a few years after the publication of the checklist, the issues of missing data, code, and model training information are still ongoing in NLP.

This project replicates the pipeline of Firebanks-Quevedo et al. (2022) in a new domain, also addressing its reproducibility issues through the publication of our novel dataset[1] and updated code[2]. Where the original workflow was based on Spanish-language texts from Latin American forestry policies, we focus on the English-language texts of Irish forestry policies. We maintained the policy domain of forestry as our implementation already introduced changes to the language and country of interest. Additionally, we further updated the pipeline with new models and accommodations for imbalanced datasets to improve performance. Our final results validate the original pipeline by demonstrating its transferability across linguistic and geopolitical contexts, as well as contribute a new resource for NLP and policy analysis through the release of our code and 1.4k sentence dataset labelled for binary and multiclass incentive classification.

2 Related Work

NLP has become a familiar tool for political and social scientists working with diverse and ever-growing text datasets for diverse and ever-growing applications. For news analysis, classifiers have been trained to identify different topics or agendas contained in articles, even so far as to identify the use of different propaganda techniques (Yoosuf and Yang, 2019; Terechshenko et al., 2020; Nelson et al., 2021). Studying political parties and their messaging, classification and topic clustering have provided novel insights into the speeches of politicians and their electoral manifestos (Glavaš et al., 2017; Wilkerson and Casas, 2017; Rheault and Cochrane, 2020). Considering public opinion, Hagen et al. (2015) used topic modelling to explore policy suggestions from public petitions, while Terechshenko et al. (2020) performed sentiment analysis on tweets about an electoral candidate. Across these various political media, NLP has successfully helped scientists to synthesise new information and extract hidden patterns from the often-unstructured text data.

Progressing more into the subdomain of public administration and policy, Żółkowski et al. (2022) applied topic modelling and clustering to explore how EU countries were framing their climate policies, and Ningpeng et al. (2024) performed topic mining and text parsing on Chinese financial policy documents. Brandt (2019) additionally examined the restoration policies of three East African countries through paragraph topic classification. While these projects provided insight into the priorities of policies, their outcomes appear to be aimed more at political scientists than specifically policy analysts. In contrast, the sentence-level incentive classification of Latin American forestry policies in Firebanks-Quevedo et al. (2022) was explicitly aimed at aiding the task of policy analysis, though the inaccessibility of the models and dataset complicated its impact. Finally, Sewerin et al. (2023) have created a policy design annotations (POLIANNA) dataset, consisting of labelled text spans from EU climate policies and legal documents to provide structured data for future policy analysis tools. In summary, the application of NLP methods to analyse policy documents has largely followed the trend of topic modelling, clustering, and classification used on other political texts, but this is beginning to shift towards the use of sentence classification on policy documents to aid in the task of policy analysis.

3 Methodology

We gathered our policies by scraping them from the website of the Irish government, processed the PDFs into sentences, cultivated a dataset through both manual and human-in-the-loop (HITL) labelling processes, then trained and evaluated both the binary and multiclass classifiers. Figure 1 demonstrates this process.

3.1 Data Collection and Preprocessing

We began looking for Irish forestry policies in a repository of policies on the website of the Irish government[3]. At the time of our search, this page was still in development. The structure of the page was a list of the Irish government's departments and major initiatives wherein which a user, upon selecting a title in the list, would be taken to that topic's page. Then, on that topic's page, a user

[1] https://huggingface.co/datasets/mawaskow/irish_forestry_incentives

[2] https://github.com/mawaskow/policy-classifier

[3] https://www.gov.ie/policies/

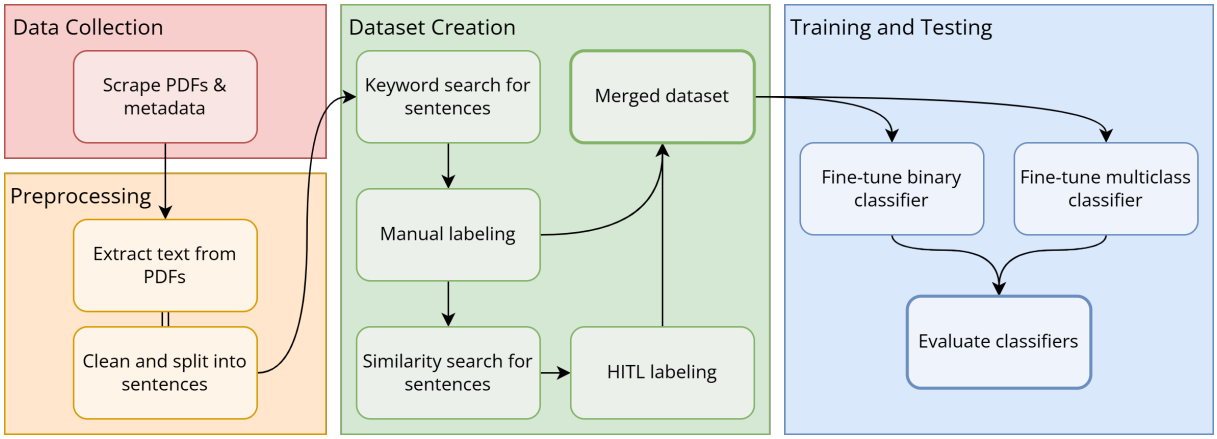

Figure 1: The incentive classifier pipeline

could find a link to a page where they could finally search that topic's publications, including policies. For a human, trying to find all the relevant policies across all the topics pages would be tedious and time-consuming, so we built a web crawler to help us scrape the data we needed.

We used Python's Scrapy[4] library to create a new "spider" or web crawler to extract the Uniform Resource Locators (URLs) for all the topic publication search pages. With this list, we constructed search query strings using keywords[5] for policy inclusion based on those used by Firebanks-Quevedo et al. (2022). Our crawler then visited all the search result pages and parsed through the queried policies to gather the policy metadata and document URLs, using exclusion keywords based on those in the original pipeline to ignore irrelevant policies and documents. In all, our scraper collected 138 relevant forestry policy documents across all the topic policy lists. A breakdown of the policies by source department can be found in Appendix A.

Once we had the collection of PDFs, we updated some of the preprocessing scripts of Firebanks-Quevedo et al. (2022) to retrieve the text from the PDFs using Python's PyPDF[6] library, which is based on text extraction instead of Optical Character Recognition (OCR). We then cleaned the texts of excess tags, spaces, and URLs and split them into sentences for labelling using a sentence tokeniser from Python's NLTK[7] library.

[4] https://www.scrapy.org/

[5] https://github.com/mawaskow/policy-classifier/blob/main/policy_scraping/policy_scraping/keywords/keywords_forestry.json

[6] https://pypdf.readthedocs.io/

[7] https://www.nltk.org/

3.2 Dataset Creation

We follow the dataset creation steps of Firebanks-Quevedo et al. (2022), first hand-labelling a subset of the collected sentences, then using those labelled sentences as the basis for a HITL process to automatically label remaining sentences. One labeller completed both the manual and HITL steps, then a policy researcher annotated a stratified subset of the data in order to validate the labelling.

We classified sentences into one of six incentive classes, or as a non-incentive, as defined in the reference pipeline:

- **Credit**: Loans, insurance
- **Direct payment**: Cash, grants
- **Fine**: Penalty payment
- **Supplies**: Material support, equipment
- **Tax deduction**: Reduced tax liability
- **Technical assistance**: Training, experts

Firebanks-Quevedo et al. (2022) had noted that their pipeline was unable to distinguish between intentions, plans, or general mentions of incentives in the sentences. In an attempt to address this, we labelled sentences that simply mentioned incentive keywords like *grant* or *loan* as a non-incentive, while sentences which actually declared the creation or implementation of incentives were classified by incentive type.

3.2.1 Manual Labelling

For our first pass at labelling the sentences, we performed a keyword search of incentive-related

substrings, then manually assigned labels to a subset of the possible incentive sentences. The hand-labelling resulted in 965 sentences across the seven classes.

3.2.2 HITL Labelling

Firebanks-Quevedo et al. (2022) conducted their HITL labelling by using five sentences for each incentive class as queries to perform a similarity search in the remaining sentences. Of the sentences returned, the original pipeline only kept sentences that occurred across all five queries for each class. As our data was reflective of one country instead of five, the size of our dataset required that we lower the inclusion criteria to occurrence within at least four of the five queries for each label. Otherwise, we followed the same steps and produced an additional 626 pre-labelled sentences across the seven classes, subsequently validated by our labeller.

3.2.3 Novel Dataset

Our final dataset of the incentive sentences was the result of merging the manually labelled and HITL-labelled datasets. In order to account for duplicate sentences across the two collections, we grouped the sentences with Levenshtein or edit distances above the hand-tuned threshold of 0.9 and removed redundant entries. The final dataset consisted of 1419 labelled, filtered sentences containing both binary and multiclass labels.

The dataset overall had a notable but expected skew towards non-incentives, with incentives making up only 18.5% of the dataset. The classes of incentives were also unbalanced themselves, with Supplies at 30.1% of all incentive sentences, Technical assistance at 27.9%, Direct payment at 23.0%, then Fine at 8.6%, Credit at 7.1%, and Tax deduction at 3.3%.

For validation of our labelling, we brought in a policy researcher to label a stratified subsample of our final dataset. They annotated 10.7% of the sentences, including 24.5% of the incentives. To evaluate our agreement, we calculated the Cohen's kappa score of the resulting binary and multiclass datasets (Cohen, 1960). The agreement of the binary labelling, computed across the entire dataset, was 0.631, considered "substantial agreement." For the multiclass labelling, we evaluated agreement across the subset of examples where both annotators labelled the sentence as some kind of incentive, resulting in a score of 0.859, "near perfect agreement." Further information about validator agreement can be found in Appendix B.

Parameter	Binary	Multiclass
Epochs	5	15
Batch Size	16	16
Learning Rate	2E-5	2E-5
Weight Decay	0.01	0.01
Optimiser	AdamW	AdamW

Table 1: Hyperparameters for binary and multiclass model training

3.3 Training and Testing

The next step in the pipeline is the fine-tuning of models on each of the binary and multiclass versions of the dataset. All training was conducted on a CUDA-enabled NVIDIA GeForce RTX 3080 Laptop GPU with 16 GB VRAM, appropriate for small to medium size models.

To produce the binary dataset, we used all 1419 sentences, keeping the non-incentive label and replacing all incentive class labels with "Incentive." The multiclass dataset consisted of the 263 incentive sentences from the final dataset with no additional alterations.

Due to the small size and significant class imbalances of the datasets, we chose to train and evaluate our classifiers across ten random train-dev-test splits (60/20/20) of each of the binary and multiclass datasets, stratified to maintain label proportions. To ensure reproducibility and controlled randomness, each split was generated using a distinct random seed (ranging from 0 to 9). We averaged our final metrics across all ten runs to account for variations in performance across different splits, providing a more robust estimate of model performance.

3.3.1 Model Selection

We first established the baseline for our replication by using the same sentence-transformer (Reimers and Gurevych, 2019) model and hyperparameters as the original pipeline for our binary and multiclass classification. The reference model, sentence-transformers/paraphrase-xlm-r-multilingual-v1 (XLM-R), is a multilingual transformer with 278M parameters (Conneau et al., 2020).

Following the first model, we explored whether newer or more efficient models could improve performance. We tried another multilingual SBERT model with the same number

Model		Binary	Multiclass
XLM-R	d	87.7 ± 1.4	94.1 ± 4.2
	l	87.5 ± 1.9	94.3 ± 3.5
	o	87.5 ± 1.3	94.7 ± 3.7
MPNet	d	**88.3 ± 1.5**	95.7 ± 2.5
	l	88.0 ± 2.3	96.2 ± 2.5
	o	88.2 ± 1.0	**96.8 ± 2.1**
GTE	d	87.1 ± 1.1	95.6 ± 3.2
	l	87.2 ± 1.4	95.3 ± 2.7
	o	87.7 ± 1.2	96.3 ± 2.7
E5	d	87.4 ± 1.6	95.7 ± 2.9
	l	86.8 ± 1.7	95.8 ± 2.0
	o	86.5 ± 2.2	94.9 ± 3.3

Table 2: Overall F1 scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Non-Incentive	Incentive
XLM-R	d	92.5 ± 1.0	66.7 ± 3.6
	l	92.4 ± 1.2	66.1 ± 4.9
	o	92.7 ± 0.8	64.7 ± 3.8
MPNet	d	93.0 ± 0.9	**68.1 ± 4.4**
	l	92.8 ± 1.4	67.0 ± 6.1
	o	**93.2 ± 0.6**	66.6 ± 2.9
GTE	d	92.2 ± 0.8	64.8 ± 2.5
	l	92.4 ± 0.9	64.9 ± 3.9
	o	92.7 ± 0.8	66.0 ± 3.5
E5	d	92.5 ± 1.0	65.3 ± 4.3
	l	92.0 ± 1.1	63.9 ± 4.3
	o	92.1 ± 1.3	62.1 ± 6.2

Table 3: Label-specific F1 scores for the binary dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

of parameters, sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (MPNet), which offers improved context capturing through its blending of permutation language modelling (PLM) with masked language modelling (MLM) (Song et al., 2020). We additionally tried two more recent, lightweight (109M parameters), English-only models, thenlper/gte-base (GTE) (Wang et al., 2022) and intfloat/e5-base-v2 (E5) (Li et al., 2023), to examine the possibility of pipeline deployment in environments with fewer computational resources–common in the political and social sciences. To ensure controlled comparisons across models, we used the same hyperparameters for all experiments as shown in Table 1.

Beyond updating the models, we also attempted to address the class imbalances of the dataset through two common strategies. In our first approach, we weighted the cross entropy loss, modifying the loss function to more severely penalise misclassification of the underrepresented classes during training; our class weights were inversely proportional to the class frequencies. Our other approach was to automatically oversample the minority classes with Imbalanced-learn's[8] RandomOverSampler, reinforcing the model's exposure to the underrepresented classes by balancing their distribution in the training data of each split.

3.3.2 Classification

Firebanks-Quevedo et al. (2022) determined that their pipeline performed best when they used the fine-tuned models to generate embeddings of the dataset, then sent those embeddings to a support vector machine (SVM) for classification (Gunn, 1998). Before we conducted the classification this way, we explored the inference capabilities of the fine-tuned transformer heads but found that performance on our small, imbalanced dataset was consistently higher when using the external classifier. For that reason and to further support our replication objective, all results reported are from the embedding generation and SVM classification method.

[8] https://imbalanced-learn.org/stable/

4 Results

We examined the F1 scores for the SVM classification of each model's sentence embeddings, averaged across the ten splits of the binary and multiclass datasets.

The mean F1 score and standard deviation for all models across the default, weighted loss, and oversampling training runs can be found in Table 2. MPNet outperformed the XLM-R model used in the reference pipeline and both other models, reporting a best average F1 score of 88.3 for binary classification and 96.8 for multiclass classification.

Overall, for binary classification, MPNet's embeddings achieved the highest average F1 scores, reporting a best value of 88.3. MPNet's performance was then followed by the XLM-R model, with the GTE and E5 models performing similarly but slightly worse than XLM-R. For multiclass classification, the MPNet again achieved the highest average F1 scores, this time reporting a best value of 96.8. After the MPNet, the rankings followed as

Model		Credit	Direct Payment	Fine	Supplies	Tax Deduction	Technical Assistance
XLM-R	d	97.1 ± 5.7	93.3 ± 5.1	93.3 ± 7.9	95.7 ± 4.0	81.7 ± 32.0	93.4 ± 4.8
	l	97.1 ± 5.7	92.1 ± 6.1	91.7 ± 9.6	96.5 ± 3.9	81.7 ± 32.0	94.7 ± 3.7
	o	**98.6 ± 4.3**	94.3 ± 5.4	94.4 ± 8.0	96.2 ± 3.7	78.3 ± 31.7	93.5 ± 3.7
MPNet	d	94.3 ± 7.0	93.5 ± 4.2	97.5 ± 7.5	97.1 ± 3.1	95.0 ± 15.0	95.7 ± 2.9
	l	93.2 ± 8.8	94.0 ± 3.9	97.8 ± 4.4	**98.4 ± 2.1**	96.7 ± 10.0	95.8 ± 4.2
	o	94.6 ± 8.6	**95.5 ± 2.8**	98.9 ± 3.3	97.5 ± 2.3	96.7 ± 10.0	**96.9 ± 2.9**
GTE	d	94.6 ± 6.7	93.7 ± 4.6	**100.0 ± 0.0**	96.5 ± 2.1	90.0 ± 30.0	95.3 ± 3.7
	l	89.8 ± 10.1	94.4 ± 3.5	**100.0 ± 0.0**	97.2 ± 2.6	90.0 ± 30.0	94.3 ± 3.6
	o	97.1 ± 5.7	95.3 ± 4.1	97.5 ± 7.5	97.2 ± 2.6	85.0 ± 32.0	96.4 ± 2.3
E5	d	94.6 ± 6.7	94.5 ± 5.0	97.8 ± 4.4	97.2 ± 2.5	96.7 ± 10.0	94.5 ± 4.4
	l	93.2 ± 6.9	94.6 ± 3.1	**100.0 ± 0.0**	96.9 ± 3.4	**100.0 ± 0.0**	94.7 ± 2.6
	o	95.7 ± 6.5	93.2 ± 4.7	97.5 ± 7.5	96.5 ± 3.1	95.0 ± 15.0	93.6 ± 4.4

Table 4: Label-specific F1 scores for the multiclass dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

GTE, then E5, then the XLM-R.

Tables 3 and 4 include the F1 scores for each classification label in the binary and multiclass datasets, respectively. Our best binary model's embeddings achieved average F1 scores of 93.0 and 68.1 for the classes of Non-incentive and Incentive, and our best multiclass classification model achieved average F1 scores of 94.6 for Credit, 95.5 for Direct payment, 98.9 for Fine, 97.5 for Supplies, 96.7 for Tax deduction, and 96.9 for Technical assistance.

We additionally report the overall validation F1 scores of our models in Appendix C as recommended by Dodge et al. (2019), as well as the overall and label-specific precision and recall scores of our models in Appendix D.

4.1 Qualitative Analysis

Examples of accurately and inaccurately classified binary incentive sentences can be found in Table 5. These sentences, embedded by MPNet and classified via SVM, highlight the semantic challenges of incentive sentence labelling and classification.

In the successfully identified examples, the label of the incentive sentence is clear due to its incentive-related keywords of *financial support*, *payments*, and *beneficiaries*, numerical amounts, and use of *will* to denote action. While the non-incentive sentence is a bit ambiguous due to the keywords *taxation* and *incentivising*, the use of *committed* demonstrates that this is a statement of intention or policy aspiration instead of a properly declared incentive, likely leading to its correct classification.

In the unsuccessful classification examples, the incentive sentence states the requirement in a scheme to apply financial penalties to herdowners but was marked as a non-incentive. Despite the use of keywords, this type of incentive– fine– may be better understood as a disincentive, so its lack of mention of rewards as found in other incentive sentences may have contributed to its misclassification. The non-incentive sentence mistakenly classified as an incentive did mention *direct payments*, but was in reality a description of a mechanism currently in place. The model likely picked up on the incentive keywords but missed the overall context of being a factual statement of existing supports rather than being an actionable incentive.

5 Discussion

Following the identification of incentive sentences through binary classification, the multiclass classification of incentive type is able to achieve high performance. In the binary context, we faced a challenge also noted by Firebanks-Quevedo et al. (2022), that it is difficult to distinguish incentive declarations from non-incentive sentences that mention incentives or state intentions to create them. At the label-specific level, the performance of incentive classification was worse than non-incentive classification, assumedly due to the imbalance of the dataset.

Regarding the different training methods, in the binary classification setting, the loss and oversampling methods did not appear to improve on default

model performance, sometimes worsening it. In the multiclass classification setting, however, XLM-R, MPNet, and GTE models all benefited from the introduction of oversampling to the training.

The multiclass labels of Credit and Tax deduction received notably higher standard deviations across models and methods, though in the case of Tax deduction, this standard deviation was reduced through updating the embedding model to MPNet as well as by adding weighted cross entropy loss or oversampling in the training process. The high standard deviations were likely due to these being the most under-represented incentive classes in the dataset, resulting in high variability of the success of their classification across training splits. The challenge of correctly classifying the Tax deduction label was also interestingly consistent with the results of Firebanks-Quevedo et al. (2022).

As our best-performing model for generating embeddings was a multilingual MPNet sentence transformer, we share in the hope of Firebanks-Quevedo et al. (2022) that the models from our pipeline can be used to classify the sentences of policies in another language via cross-lingual transfer learning, removing the need to construct a whole new dataset for fine-tuning.

While the outcomes of this research primarily serve the development and evaluation of NLP applications for incentive sentence classification in policy texts, the practical impact on policy analysts remains indirect at this stage. As most policy analysts are not comfortable with building their own NLP pipeline implementations, the immediate application of this dataset and models is limited. However, we envision future work that translates these models into accessible tools, namely an interface where policy analysts can input search terms and websites of policy repositories to automatically retrieve relevant documents, or where analysts can upload policy documents to extract incentive sentences using the classification pipeline– first identifying incentives with our binary classifier, then categorising them with our muticlass classifier. We hope to soon make our pipeline accessible in this way to bridge the gap to impact real-world policy analysis, with the ultimate goal of improving the policymaking process and preventing policy failure.

6 Conclusion

This paper applied an existing policy incentive classification pipeline to a new geopolitical and linguistic context, demonstrating the transferability of the reference pipeline and creating a novel dataset of Irish forestry policy sentences labelled for incentive classification. The binary and multiclass classification of sentence embeddings produced by our best models achieved similar performance to the original Firebanks-Quevedo et al. (2022) pipeline's results on their own policy dataset, serving as a validation of their methodology.

We prioritised reproducibility in this replication, and encourage researchers to implement this or similar pipelines for policy incentive classification in more domains across languages, political contexts, and policy areas. Additionally, our dataset and training information is available for anyone who wants to fine-tune their own multilingual incentive classification models on an existing dataset for transfer learning into new contexts.

We hope that with more progress in the area of automatic policy incentive or instrument classification, this work can help streamline the task of policy analysis to enable robust recommendations of policy solutions, ultimately working towards the creation of more successful policies.

Limitations

Our pipeline did encounter limitations, some of which were shared with Firebanks-Quevedo et al. (2022) and some of which were unique to this implementation. We encountered the same challenge that Firebanks-Quevedo et al. (2022) did concerning the ambiguity of incentive sentences across intentions, plans, and mentions, resulting in suboptimal incentive identification performance in the binary classification context. Further examples of ambiguous sentences are presented in Appendix E.

Specific to this pipeline, in our labelling process we found that there were no incentive sentences about providing direct material support or equipment, so we adapted the Supplies definition to include grants which were exclusively for the purchase of materials or equipment. Additionally, when it came to labelling sentences about carbon taxes, we were conflicted between labelling it a Tax deduction since it is a tax mechanism, or including it with Fine since it can be considered a penalty or disincentive; we decided to proceed with the latter option. Additionally, the training and testing of our

Successes	True Label
Financial support towards the professional costs, such as legal, taxation and advisory for older farmers will contribute 50% of such vouched costs, to a maximum payment of €1,500 per beneficiary.	Incentive
We are committed to further developing a taxation framework, which plays its full part in incentivising, along with other available policy levers, the necessary actions to reduce our emissions.	Non-incentive
Errors	**True Label**
Where it can be established that such ineligible features/areas existed in previous years, there is a requirement to reduce the area and apply the relevant financial reduction and/or penalty to the herdowner.	Incentive
EU CAP direct payments provide vital income support for farmers, and act as an important cushion against commodity price volatility.	Non-incentive

Table 5: Sentences from the binary dataset correctly and incorrectly classified with our best-performing model

embedding models and classifier were limited by the small size and class imbalances of the dataset.

We also noted that the pipeline may benefit from span extraction and classification rather than simple sentence classification, especially in cases where there are several incentives contained in a single sentence. We are now experimenting with the POLIANNA dataset for extracting spans to use as features in the sentence classification, enabling multilabel classification for incentives as well (Sewerin et al., 2023).

Ethics

The data for this project consists of sentences from public policy documents, none of which contain private or personal information. Bias may be introduced to the dataset and resulting models by the aforementioned single annotator and annotation decisions in ambiguous cases, as well as by the small size and class imbalance of the dataset. While our pipeline is primarily designed and intended for policy analysis, it could be manipulated to create biased classifiers that mislabel incentives in order to fit a certain agenda, potentially over-representing or under-representing different incentive classes in order to sway the downstream choice of policy solutions. As such, it is important for future implementations of the pipeline to maintain openness and transparency in the construction of their datasets and training of their models.

Acknowledgments

Thank you to Dr. Yifan Wang for annotating a subset of our data for validation. This work was funded by Research Ireland as part of Grant Number SFI/12/RC/2289_P2 Insight_2, the Insight Research Ireland Centre for Data Analytics.

References

Bertrand Badie, Dirk Berg-Schlosser, and Leonardo Morlino. 2011. *International Encyclopedia of Political Science*, volume 1. SAGE Publications, Inc., Thousand Oaks, California.

John Brandt. 2019. Text mining policy: Classifying forest and landscape restoration policy agenda with neural information retrieval. In *KDD Fragile Earth workshop (FEED 2019)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 2185–2194.

Daniel Firebanks-Quevedo, Jordi Planas, Kathleen Buckingham, Cristina Taylor, David Silva, Galina Naydenova, and René Zamora-Cristales. 2022. Using machine learning to identify incentives in forestry policy: Towards a new paradigm in policy analysis. *Forest Policy and Economics*, 134:102624.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 42–46, Vancouver, Canada. Association for Computational Linguistics.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Steve R. Gunn. 1998. Support vector machines for classification and regression. Project report, University of Southampton. Address: Southampton, U.K.

Loni Hagen, Özlem Uzuner, Christopher Kotfila, Teresa M. Harrison, and Dan Lamanna. 2015. Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In *2015 48th Hawaii International Conference on System Sciences*, pages 2134–2143.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What have we learned from the checklist? *Findings of the Association for Computational Linguistics: ACL 2023*, page 12789–12811.

Laura K. Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1):202–237.

Jiang Ningpeng, Han Tian, Wang Haibo, Xu Ruzhi, and Ma Shiyu. 2024. A study on structured text parsing for policies based on BERTopic. In *2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, volume 6, pages 16–22.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ludovic Rheault and Christopher Cochrane. 2020. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.

Sebastian Sewerin, Lynn H. Kaack, Joel Küttel, Fride Sigurdsson, Onerva Martikainen, Alisha Esshaki, and Fabian Hafner. 2023. Towards understanding policy design through text-as-data approaches: The policy design annotations (polianna) dataset. *Scientific Data*, 10(896).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zhanna Terechshenko, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2020. A comparison of methods in political science text classification: Transfer learning language models for politics. *Other Information Systems & eBusiness eJournal*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

John Wilkerson and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20(Volume 20, 2017):529–544.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.

Artur Żółkowski, Mateusz Krzyziński, Piotr Wilczyński, Stanisław Giziński, Emilia Wiśnios, Bartosz Pieliński, Julian Sienkiewicz, and Przemysław Biecek. 2022. Climate policy tracker: Pipeline for automated analysis of public climate policies. In *Tackling Climate Change with Machine Learning Workshop*.

A Policy Sources

Table 6 shows the departments of origin for the policies in the dataset at the time of collection. Since the collection however, two of the departments have been renamed and their websites and URLs restructured. In order to address this issue of changing PDF sources and addresses, we provide a ZIP file[9] of the policies in our repository.

[9] https://github.com/mawaskow/policy-classifier/blob/main/policy_scraping/policy_scraping/outputs/forestry/full.zip

# Policies	Department
45	Agriculture, Food and the Marine
45	Rural and Community Development
44	The Environment, Climate and Communications
4	Housing, Local Government and Heritage

Table 6: The number of policies from each Irish government department represented in the dataset

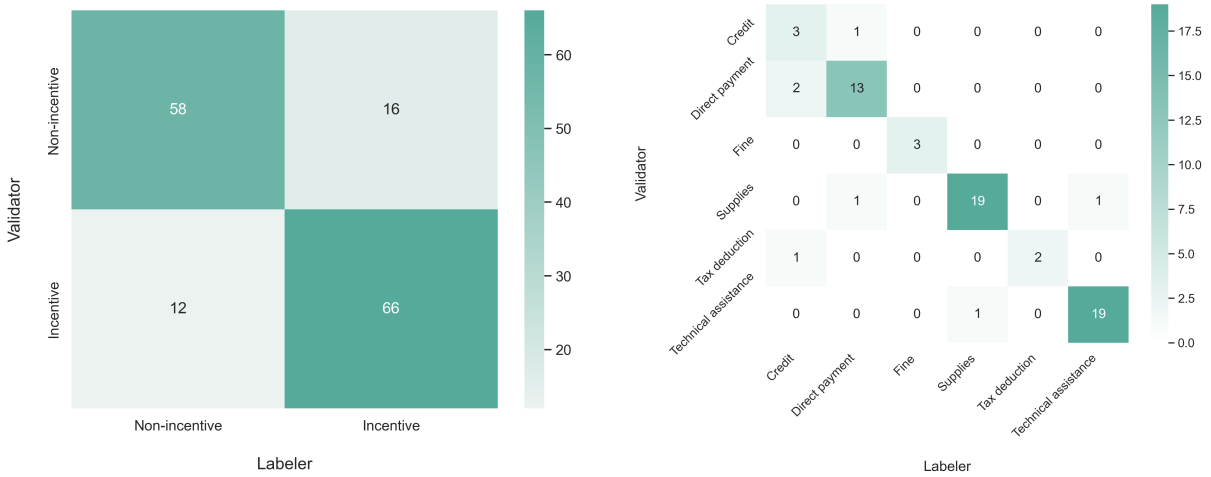

Figure 2: Confusion matrices for the validation sample (left: binary, right: multiclass)

B External Validation

For the external validation of our dataset, we provide the confusion matrices of our subset's labelling in Figure 2 for binary and multiclass classification. Additional information about annotation guidelines can be found on the dataset page.

C Validation Performance

Table 7 reflects performance using the transformer head for classification rather than the SVM classifier used in our main pipeline, resulting in the lower scores observed here. Despite this, the results confirm that MPNet consistently outperforms the other models in incentive sentence embedding for both binary and multiclass classification.

D Precision and Recall

In addition to reporting the overall and label-specific F1 scores of our models' binary and multiclass classification experiments in Tables 2, 3, and 4, we report their precision in Tables 8, 10, and 12, and recall in Tables 9, 11, and 13.

E Ambiguous Cases

Table 14 provides further examples of ambiguous policy sentences.

Model		Binary	Multiclass
XLM-R	d	88.4 ± 2.7	87.2 ± 3.5
	l	88.0 ± 2.6	89.3 ± 3.0
	o	89.1 ± 2.2	91.5 ± 3.7
MPNet	d	**89.8 ± 2.7**	89.0 ± 2.9
	l	89.2 ± 2.2	91.7 ± 3.0
	o	89.1 ± 2.0	**91.8 ± 3.1**
GTE	d	88.4 ± 2.5	73.5 ± 7.3
	l	88.2 ± 2.6	82.9 ± 5.1
	o	89.4 ± 2.3	90.5 ± 2.8
E5	d	88.0 ± 2.8	83.6 ± 3.1
	l	88.1 ± 2.2	88.1 ± 5.0
	o	89.6 ± 2.1	90.4 ± 3.6

Table 7: Overall validation F1 scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Binary	Multiclass
XLM-R	d	87.7 ± 1.4	94.6 ± 3.8
	l	87.5 ± 1.9	95.0 ± 3.0
	o	87.4 ± 1.4	95.3 ± 3.7
MPNet	d	**88.3 ± 1.6**	96.3 ± 2.1
	l	87.9 ± 2.2	96.6 ± 2.3
	o	88.1 ± 1.0	**97.0 ± 1.9**
GTE	d	87.1 ± 1.1	95.8 ± 3.1
	l	87.2 ± 1.4	95.7 ± 2.6
	o	87.6 ± 1.3	96.7 ± 2.6
E5	d	87.3 ± 1.6	96.0 ± 2.8
	l	86.7 ± 1.7	96.1 ± 1.9
	o	86.3 ± 2.3	95.5 ± 2.9

Table 8: Overall precision scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Binary	Multiclass
XLM-R	d	87.8 ± 1.5	94.2 ± 4.2
	l	87.6 ± 2.0	94.3 ± 3.5
	o	87.9 ± 1.3	94.7 ± 3.5
MPNet	d	88.5 ± 1.5	95.7 ± 2.5
	l	88.2 ± 2.3	96.2 ± 2.5
	o	**88.7 ± 1.0**	**96.8 ± 2.1**
GTE	d	87.3 ± 1.2	95.7 ± 3.0
	l	87.5 ± 1.4	95.5 ± 2.6
	o	87.9 ± 1.2	96.4 ± 2.6
E5	d	87.6 ± 1.6	95.7 ± 2.9
	l	86.9 ± 1.8	95.8 ± 2.0
	o	86.9 ± 2.1	94.9 ± 3.4

Table 9: Overall recall scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Non-Incentive	Incentive
XLM-R	d	92.1 ± 0.9	68.5 ± 5.7
	l	92.0 ± 1.1	67.9 ± 6.5
	o	91.0 ± 1.0	71.4 ± 5.0
MPNet	d	**92.3 ± 1.1**	70.6 ± 4.8
	l	92.0 ± 1.4	70.2 ± 7.0
	o	91.4 ± 0.7	**74.0 ± 3.6**
GTE	d	91.6 ± 0.8	67.6 ± 5.7
	l	91.5 ± 1.0	68.2 ± 5.1
	o	91.7 ± 1.1	69.8 ± 4.9
E5	d	91.5 ± 1.1	68.8 ± 5.2
	l	91.3 ± 0.9	66.4 ± 6.1
	o	90.6 ± 1.4	67.7 ± 6.9

Table 10: Label-specific precision scores for the binary dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Non-Incentive	Incentive
XLM-R	d	93.0 ± 1.7	65.3 ± 4.2
	l	92.9 ± 1.8	64.7 ± 5.0
	o	94.5 ± 1.4	59.4 ± 4.9
MPNet	d	93.6 ± 1.4	**66.0 ± 5.5**
	l	93.6 ± 2.1	64.5 ± 6.7
	o	**95.1 ± 0.9**	60.8 ± 3.5
GTE	d	92.9 ± 1.8	62.6 ± 4.1
	l	93.2 ± 1.6	62.3 ± 5.1
	o	93.6 ± 1.5	63.0 ± 5.4
E5	d	93.5 ± 1.4	62.3 ± 5.1
	l	92.7 ± 1.7	61.7 ± 4.0
	o	93.6 ± 1.6	57.5 ± 6.8

Table 11: Label-specific recall scores for the binary dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Credit	Direct Payment	Fine	Supplies	Tax Deduction	Technical Assistance
XLM-R	d	**100.0 ± 0.0**	93.4 ± 5.0	98.3 ± 5.0	96.1 ± 6.1	78.3 ± 35.0	92.5 ± 5.5
	l	**100.0 ± 0.0**	91.5 ± 7.0	**100.0 ± 0.0**	96.9 ± 4.0	78.3 ± 35.0	93.8 ± 5.4
	o	**100.0 ± 0.0**	**95.9 ± 4.2**	98.3 ± 5.0	96.9 ± 4.1	73.3 ± 35.1	92.2 ± 6.4
MPNet	d	**100.0 ± 0.0**	91.9 ± 6.5	**100.0 ± 0.0**	**98.2 ± 2.8**	93.3 ± 20.0	95.9 ± 5.7
	l	97.5 ± 7.5	91.9 ± 6.7	**100.0 ± 0.0**	**98.2 ± 2.8**	95.0 ± 15.0	97.4 ± 3.3
	o	97.5 ± 7.5	95.4 ± 5.0	**100.0 ± 0.0**	97.0 ± 3.0	95.0 ± 15.0	**97.5 ± 3.1**
GTE	d	98.0 ± 6.0	91.2 ± 7.3	**100.0 ± 0.0**	97.6 ± 3.8	90.0 ± 30.0	96.1 ± 4.3
	l	98.0 ± 6.0	92.7 ± 6.9	**100.0 ± 0.0**	97.6 ± 3.8	90.0 ± 30.0	94.5 ± 5.6
	o	**100.0 ± 0.0**	95.9 ± 4.1	**100.0 ± 0.0**	96.9 ± 3.1	83.3 ± 34.2	95.8 ± 4.6
E5	d	98.0 ± 6.0	95.1 ± 5.5	**100.0 ± 0.0**	97.6 ± 3.9	95.5 ± 15.0	93.1 ± 5.8
	l	98.0 ± 6.0	94.5 ± 5.0	**100.0 ± 0.0**	97.0 ± 4.0	**100.0 ± 0.0**	94.3 ± 4.3
	o	**100.0 ± 0.0**	94.5 ± 6.1	**100.0 ± 0.0**	97.5 ± 3.0	93.3 ± 20.0	91.7 ± 6.8

Table 12: Label-specific precision scores for the multiclass dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Credit	Direct Payment	Fine	Supplies	Tax Deduction	Technical Assistance
XLM-R	d	95.0 ± 10.0	93.3 ± 6.2	90.0 ± 13.4	95.6 ± 4.9	90.0 ± 30.0	94.7 ± 7.2
	l	95.0 ± 10.0	93.3 ± 9.0	86.0 ± 15.6	96.3 ± 5.0	90.0 ± 30.0	96.0 ± 5.3
	o	**97.5 ± 7.5**	93.3 ± 9.0	92.0 ± 13.3	95.6 ± 4.9	90.0 ± 30.0	95.3 ± 5.2
MPNet	d	90.0 ± 12.2	95.8 ± 6.7	96.0 ± 12.0	96.3 ± 5.0	**100.0 ± 0.0**	96.0 ± 5.3
	l	90.0 ± 12.2	**96.7 ± 4.1**	96.0 ± 8.0	**98.8 ± 2.5**	**100.0 ± 0.0**	94.7 ± 7.2
	o	92.5 ± 11.5	95.8 ± 4.2	98.0 ± 6.0	98.1 ± 2.9	**100.0 ± 0.0**	96.7 ± 5.4
GTE	d	92.5 ± 11.5	**96.7 ± 4.1**	**100.0 ± 0.0**	95.6 ± 2.9	90.0 ± 30.0	94.7 ± 5.0
	l	85.0 ± 16.6	**96.7 ± 4.1**	**100.0 ± 0.0**	96.9 ± 3.1	90.0 ± 30.0	94.7 ± 5.8
	o	95.0 ± 10.0	95.0 ± 6.7	96.0 ± 12.0	97.5 ± 3.1	90.0 ± 30.0	**97.3 ± 4.4**
E5	d	92.5 ± 11.5	94.2 ± 6.5	96.0 ± 8.0	96.9 ± 3.1	**100.0 ± 0.0**	96.0 ± 4.4
	l	90.0 ± 12.2	95.0 ± 4.1	**100.0 ± 0.0**	96.9 ± 4.2	**100.0 ± 0.0**	95.3 ± 4.3
	o	92.5 ± 11.5	92.5 ± 6.9	96.0 ± 12.0	95.6 ± 4.9	**100.0 ± 0.0**	96.0 ± 5.3

Table 13: Label-specific recall scores for the multiclass dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

The Eco-Schemes provides supports to farmers who undertake specific agricultural practices, including extensive farming, tree planting, sewing a multi-species sward, and enhancing crop diversification.
Department of Finance and DAFM to improve dissemination of information on taxation incentives including engagement with advisors, tax consultants and accountants.
LEADER may provide support rates greater than 65% in accordance with Article 73(4) (c)(ii) where investments include basic services in rural areas and infrastructure in agriculture and forestry, as determined by Member States.

Table 14: Ambiguous policy sentences, unclear in either their status as incentive or non-incentive, or in the type of incentive they are.